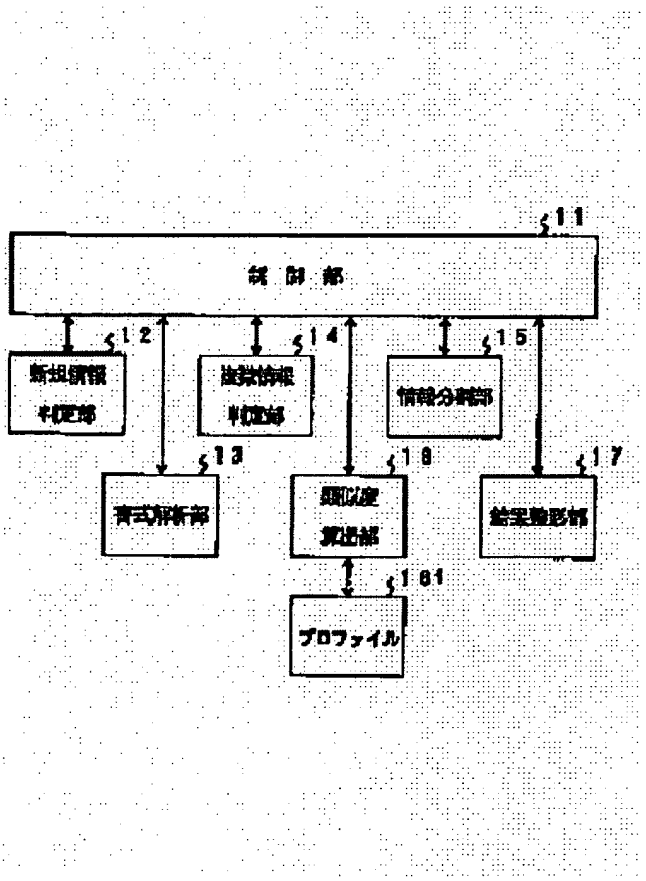
INFORMATION FILTERING DEVICE AND ITS METHOD

Patent number: JP10143541
Publication date: 1998-05-29
Inventor: SUMITA KAZUO
Applicant: TOKYO SHIBAURA ELECTRIC CO
Classification:
- international: G06F17/30
- european:
Application number: JP19970249100 19970912
Priority number(s): JP19960243785 19960913; JP19970249100 19970912

Abstract of **JP10143541**

PROBLEM TO BE SOLVED: To obtain the information filtering device narrowing and providing only information necessitated by a user by setting a document prepared and corrected irregularly to be an object.
SOLUTION: In this device, similarity between a retrieving condition registered in a profile 161 in advance and information included in a document to be a processing object is calculated by a similarity calculating part 16 and a prescribed document is selected from among plural documents according to the calculated similarity. A plural information judging part 14 judges whether the document includes plural information units and an information dividing part 15 divides a document judged to include plural information units by the part 14 by an each information unit. Then the part 16 calculates similarity to the document for each information unit included in the document. Consequently each of the information units within the document including plural information units is filter-processed without receiving any influence from surrounding information.



(19)日本国特許庁（ＪＰ）　(12)公開特許公報（A）　(11)特許出願公開番号

特開平10－143541

(43)公開日　平成10年(1998)５月29日

(51)Int.Cl.[6]	識別記号	ＦＩ		
Ｇ０６Ｆ　17/30		Ｇ０６Ｆ	15/403	３４０Ａ
			15/40	３７０Ａ
			15/401	３１０Ｄ

審査請求　未請求　請求項の数９　ＯＬ　（全 19 頁）

(21)出願番号　特願平9－249100

(22)出願日　平成９年(1997)９月12日

(31)優先権主張番号　特願平8－243785

(32)優先日　平８(1996)９月13日

(33)優先権主張国　日本（ＪＰ）

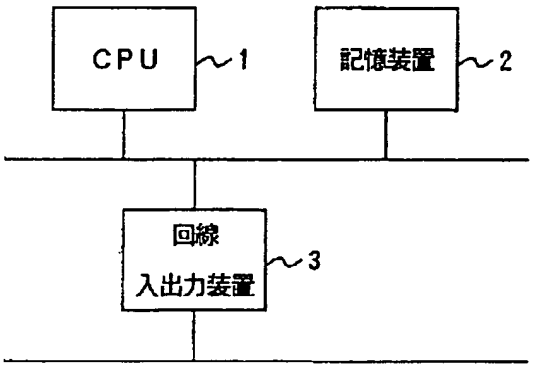
(71)出願人　000003078
株式会社東芝
神奈川県川崎市幸区堀川町72番地

(72)発明者　住田　一男
神奈川県川崎市幸区小向東芝町１番地　株式会社東芝研究開発センター内

(74)代理人　弁理士　鈴江　武彦　（外６名）

(54)【発明の名称】　情報フィルタリング装置および情報フィルタリング方法

(57)【要約】
【課題】非定期的に発生および修正される文書を対象にして、ユーザが必要とする情報のみを絞り込んでユーザに提供する情報フィルタリング装置。
【解決手段】予めプロファイル１６１に登録された検索条件と、処理対象となる文書に含まれる情報との間の類似度を類似度算出部１６が算出し、その算出した類似度にしたがって、複数の文書の中から所定の文書を選出する情報フィルタリング装置において、複数情報判定部１４は、その文書が複数の情報単位を含むか否か判定し、情報分割部１５は、複数情報判定部１４によって複数の情報単位を含むと判定された文書を情報単位ごとに分割する。そして、類似度算出部１６は、文書に対する類似度を、その文書に含まれる情報単位それぞれに算出する。したがって、複数の情報単位を含む文書内の情報単位それぞれが、回りの情報に何等影響されることなく、フィルタリング処理されることになる。



【特許請求の範囲】
【請求項1】　予め登録された検索条件と文書に含まれる情報との間の類似度を算出し、その算出した類似度にしたがって複数の文書の中から所定の文書を選出する情報フィルタリング装置において、
前記文書が複数の情報単位を含むか否か判定する判定手段と、
前記判定手段によって複数の情報単位を含むと判定された文書を情報単位ごとに分割する分割手段と、
前記分割手段によって分割された情報単位それぞれに、前記検索条件との間の類似度を算出する類似度算出手段とを具備してなることを特徴とする情報フィルタリング装置。
【請求項2】　複数の文書の中から所定の文書を選出する情報フィルタリング装置であって、階層構造をなすハイパーテキストをフィルタリング対象の文書に含む情報フィルタリング装置において、
新たな情報が発生したか否か監視すべき文書のアドレスを設定する第1の設定手段と、
前記第1の設定手段によって設定された文書を起点に下位層に位置する文書に対する監視すべき階層数を設定する第2の設定手段と、
前記第1の設定手段によって設定されたアドレスから前記第2の設定手段によって設定された階層数を対象範囲として文書を読み込み、その範囲内に新たな情報が発生したか否か判定する判定手段とを具備してなることを特徴とする情報フィルタリング装置。
【請求項3】　複数の文書の中から所定の文書を選出する情報フィルタリング装置において、
他の情報フィルタリング装置により出力されるフィルタリング結果を取り込む取り込み手段と、
この取り込み手段が取り込んだフィルタリング結果を前記複数の文書に含めてフィルタリング処理を実行するフィルタリング手段とを具備してなることを特徴とする情報フィルタリング装置。
【請求項4】　予め登録された検索条件と文書に含まれる情報との間の類似度を算出し、その算出した類似度にしたがって複数の文書の中から所定の文書を選出する情報フィルタリング方法において、
前記文書が複数の情報単位を含むか否か判定し、
複数の情報単位を含むと判定された文書を情報単位ごとに分割し、
この分割された情報単位それぞれに、前記検索条件との間の類似度を算出することを特徴とする情報フィルタリング方法。
【請求項5】　複数の文書の中から所定の文書を選出する情報フィルタリング方法であって、階層構造をなすハイパーテキストをフィルタリング対象の文書に含む情報フィルタリング方法において、
新たな情報が発生したか否か監視すべき文書のアドレスを設定し、
この設定された文書を起点に下位層に位置する文書に対する監視すべき階層数を設定し、
前記設定されたアドレスから前記設定された階層数を対象範囲として文書を読み込み、その範囲内に新たな情報が発生したか否か判定することを特徴とする情報フィルタリング方法。
【請求項6】　複数の文書の中から所定の文書を選出する情報フィルタリング方法において、
他の情報フィルタリング装置が出力するフィルタリング結果を取り込み、
この取り込んだフィルタリング結果を前記複数の文書に含めてフィルタリング処理を実行することを特徴とする情報フィルタリング方法。
【請求項7】　予め登録された検索条件と文書に含まれる情報との間の類似度を算出し、その算出した類似度にしたがって複数の文書の中から所定の文書を選出するためのプログラムであって、
前記文書が複数の情報単位を含むか否か判定し、
複数の情報単位を含むと判定された文書を情報単位ごとに分割し、
この分割された情報単位それぞれに、前記検索条件との間の類似度を算出するようにコンピュータを動作させるプログラムを記録したコンピュータ読み込み可能な記録媒体。
【請求項8】　階層構造をなすハイパーテキストを含む複数の文書の中から所定の文書を選出するためのプログラムであって、
新たな情報が発生したか否か監視すべき文書のアドレスを設定し、
この設定された文書を起点に下位層に位置する文書に対する監視すべき階層数を設定し、
前記設定されたアドレスから前記設定された階層数を対象範囲として文書を読み込み、その範囲内に新たな情報が発生したか否か判定するようにコンピュータを動作させるプログラムを記録したコンピュータ読み込み可能な記録媒体。
【請求項9】　複数の文書の中から所定の文書を選出するためのプログラムであって、
他の情報フィルタリング装置が出力するフィルタリング結果を取り込み、
この取り込んだフィルタリング結果を前記複数の文書に含めてフィルタリング処理を実行するようにコンピュータを動作させるプログラムを記録したコンピュータ読み込み可能な記録媒体。
【発明の詳細な説明】
【0001】
【発明の属する技術分野】この発明は、膨大な数のテキスト記事や文献などの文書から、新たに入力された情報であってユーザの要求・興味にあったものを選出してユ

ーザに提供する情報フィルタリング装置および情報フィルタリング方法に関する。

【0002】

【従来の技術】近年、インターネットの普及は目覚ましいものがあり、世界中に点在する計算機に格納された情報が、インターネットに接続されてさえいれば、どこからでも簡単にアクセスできるようになってきている。特に、WWW（Ｗｏｒｌｄ　Ｗｉｄｅ　Ｗｅｂ）では、ＨＴＴＰ（ＨｙｐｅｒＴｅｘｔ　Ｔｒａｎｓｆｅｒ　Ｐｒｏｔｏｃｏｌ）を用いることにより、利用者が、世界中の情報をＧＵＩ（Ｇｒａｐｈｉｃａｌ　Ｕｓｅｒ　Ｉｎｔｅｒｆａｃｅ）ベースのブラウザによって簡単にアクセスできる仕組みを提供している。

【0003】WWWでは、ある計算機上でｈｔｔｐｄと呼ばれるソフトウェアを用いる。このソフトウェアは、その計算機のデータベースに格納されているＨＴＭＬ（ＨｙｐｅｒＴｅｘｔ　Ｍａｒｋｕｐ　Ｌａｎｇｕａｇｅ）で記述したハイパーテキストファイルを、他の計算機からの要求に応じて転送するものである。インターネットに接続されている計算機は、転送を要求するハイパーテキストファイルが存在するｈｔｔｐｄに対し、ハイパーテキストファイルのアドレスを指定することによって、指定したファイルを読み込むことができる。ＨＴＭＬの記述では、ハイパーテキストファイル内のリンク情報として、前記アドレスが記述されるので、ＨＴＴＰのプロトコルにしたがったブラウザは、各ｈｔｔｐｄ支配下のハイパーテキストファイルを表示することができる。そして、音声、静止画、動画などの様々なデータを出力できるようにすることによって、マルチメディアデータを含むハイパーテキストを、ブラウザは表示することができる。

【0004】このWWWの仕組みにより、利用者は、より簡単にインターネット上の情報にアクセスできるようになり、多くの個人や企業が、Ｗｅｂページと呼ばれるハイパーテキストファイルを公開するようになってきている。

【0005】しかしながら、WWWではデータベースの管理者がおらず、個々人がそれぞれ勝手にＷｅｂページを作成および修正し、しかもその規模が膨大であるために（１９９６年度初頭における世界中で公開されているＷｅｂページは４０００万ページと推定されている）、個々の利用者が自らが必要とするＷｅｂページがどこにあるか（ＵＲＬアドレスとして何を指定すれば必要なＷｅｂページを取得できるか）を知ることが困難な状況になっている。

【0006】このため、最近では、アクセス可能なＷｅｂページを内容ベースで検索するシステムが開発され、検索を代行するようなサービスが行なわれるようになってきた。具体的には、Ｙａｈｏｏ、ＬｙｃｏｓおよびＡｌｔａｖｉｓｔａなどといったＷｅｂ検索サーバが存在する。Ｗｅｂ検索サーバでは、キーワードを指定することによって、そのキーワードを含むＷｅｂページを検索することができる。利用者は、これらＷｅｂ検索サーバを用いて必要なＷｅｂページを検索する。

【0007】しかし、このようにＷｅｂ検索サーバを用いることによってオンラインで必要な情報を容易に検索できるようになったものの、これは利用者が能動的に必要な情報を検索指示した場合にのみ得られるのであって、利用者が関心・興味をいだいている情報が新しく作成された際に利用者が検索指示を行なわなければ、たとえ重要な情報であったとしても、その利用者がその情報を知ることはない。したがって、利用者が関心・興味のある情報が発生したときに、その旨を適切な利用者に知らしめるシステムが必要である。旧来のデータベースシステムでは、このような機能をＳＤＩ（Ｓｅｌｅｃｔｉｖｅ　Ｄｉｓｓｅｍｉｎａｔｉｖｅ　Ｉｎｆｏｒｍａｔｉｏｎ）と呼んでいる。ＳＤＩでは、利用者は自らの関心・興味のある情報を選択するためのキーワードなどを個人プロファイルとしてシステムに登録しておく。そして、システムは、新しくデータが登録された際に、そのデータとキーワード（プロファイル）とを比較して、そのデータがキーワードと合致するときに、所望した情報が新たに発生した旨をプロファイルの登録利用者に知らせるものである。

【0008】しかしながら、WWWでは、Ｗｅｂページにどのような情報を記載するかは個々人の自由であるという性格をもつために、一つのＷｅｂページに複数の情報単位が記載されることは十分に考えられる。そして、互いに関連のない複数の情報単位が記載されたＷｅｂページを一つの処理単位としてプロファイルとの比較を実行した場合、必ずしも適切なフィルタリングが施される保証はない。したがって、利用者が関心・興味をもつ極めて重要な情報が一部に含まれるＷｅｂページであっても、ページ全体としてその取捨が判定された結果、選択対象とならない場合が発生するといった問題があった。

【0009】また、前述したような旧来型のデータベースでは、個々のデータはローカルな環境に存在するか、または特定のデータベース管理者が管理するものであったために、新しく情報が発生した情報と既存の情報とを区別することが容易であったが、WWWでは、個々人がＷｅｂページを独自に登録できる仕組みになっており、かつWWW全体を管理する管理者も存在しないため、新規情報と既存情報との区別が非常に困難である。さらに、Ｗｅｂページは、ハイパーテキスト構造をもち、互いに関連づけられた複数のページによって一定の情報を表現することがあるため、監視対象とするページについて新規情報の発生を検出するのみでは不十分であるといった問題があった。

【0010】さらに、WWW上のＷｅｂページなどのように非常に広範囲な範囲に対して新規発生情報を監視す

ることは、単独のシステムにおいては困難であるといった問題があった。

【0011】

【発明が解決しようとする課題】このように、従来の情報フィルタリングをたとえばWWW上のＷｅｂページなどに適用する場合においては、以下に示すような問題が存在していた。

【0012】（1）Ｗｅｂページは単一の情報からなる場合と複数の情報からなる場合があり、複数の情報からなるページの場合に、個々の情報単位ごとに分割し、その情報単位ごとにプロファイルとの比較を行なわないと、必要な情報の選択が正確にできない。

【0013】（2）大規模なシステムでない場合、全世界のページを網羅的にチェックすることは単独システムでは不可能である。一方、特定のページを指定して、そのページの情報が修正されたことを検出する監視手段を設けることで、利用者の便を図ることができる。しかしながら、Ｗｅｂページはハイパーテキストであるために、複数のページによって一定の情報を表現することがあり、前述の監視手段が一つのＷｅｂページだけしか指定できないと、そのページからリンクを張られている子供ページや孫ページが修正されても検出できない。

【0014】（3）単独の情報フィルタリング装置の処理だけでは、利用者にとって十分な範囲の新規発生情報を監視することが困難である。

【0015】この発明は、このような実情に鑑みてなされたものであり、WWWのように個々人が独自にデータを作成および修正するデータベースにおいて、新規に発生した情報（新鮮な情報）の中から、利用者の関心・興味のある情報のみを効率的に選択して通知することを可能とする情報フィルタリング装置および情報フィルタリング方法を提供することを目的とする。

【0016】

【課題を解決するための手段】第1の発明の情報フィルタリング装置は、予め登録された検索条件と文書に含まれる情報との間の類似度を算出し、その算出した類似度にしたがって複数の文書の中から所定の文書を選出する情報フィルタリング装置において、前記文書が複数の情報単位を含むか否か判定する判定手段と、前記判定手段によって複数の情報単位を含むと判定された文書を情報単位ごとに分割する分割手段と、前記分割手段によって分割された情報単位それぞれに、前記検索条件との間の類似度を算出する類似度算出手段とを具備してなることを特徴とする。

【0017】この第1の発明の情報フィルタリング装置においては、判定手段が、文書それぞれに対して、単一の内容からなるデータか複数の内容からなるデータかを判定する。そして、この判定手段によって複数の内容からなるデータと判定されたときに、分割手段が、その内容ごとにフィルタリング処理を行なうべく文書を情報単位ごとに分割する。そして、類似度算出手段は、この分割された情報単位それぞれに、検索条件との間の類似度を算出する。これにより、この第1の発明の情報フィルタリング装置では、単一の内容からなるＷｅｂページと複数の内容からなるＷｅｂページとに対し、これらを同時にフィルタリング対象とし、かつ内容に応じた高精度のフィルタリングを可能とすることができる。

【0018】また、第2の発明の情報フィルタリング装置は、複数の文書の中から所定の文書を選出する情報フィルタリング装置であって、階層構造をなすハイパーテキストをフィルタリング対象の文書に含む情報フィルタリング装置において、新たな情報が発生したか否か監視すべき文書のアドレスを設定する第1の設定手段と、前記第1の設定手段によって設定された文書を起点に下位層に位置する文書に対する監視すべき階層数を設定する第2の設定手段と、前記第1の設定手段によって設定されたアドレスから前記第2の設定手段によって設定された階層数を対象範囲として文書を読み込み、その範囲内に新たな情報が発生したか否か判定する判定手段とを具備してなることを特徴とする。

【0019】この第2の発明の情報フィルタリング装置においては、第1の設定手段が、監視すべき文書を設定し、第2の設定手段が、第1の設定手段によって設定された文書を起点とした階層数を設定する。そして、判定手段が、この第1および第2の設定手段で設定された範囲のデータを対象にフィルタリング処理を行なう。これにより、階層的なＷｅｂページを監視可能とし、指定した範囲内に新規または修正された情報があるときに、それをもれなく検知することを可能とする。

【0020】また、第3の発明の情報フィルタリング装置は、複数の文書の中から所定の文書を選出する情報フィルタリング装置において、他の情報フィルタリング装置により出力されるフィルタリング結果を取り込む取り込み手段と、この取り込み手段が取り込んだフィルタリング結果を前記複数の文書に含めてフィルタリング処理を実行するフィルタリング手段とを具備してなることを特徴とする。

【0021】この第3の発明の情報フィルタリング装置によれば、他の情報フィルタリング装置が出力したフィルタリング結果を取り込むことにより、単独の情報フィルタリング装置が監視できる以上の範囲の情報を監視することを可能にする。

【0022】

【発明の実施の形態】以下、図面を参照してこの発明の実施形態について説明する。

【0023】（第1実施形態）まず、この発明の第1の実施形態について説明する。図1に本実施形態の情報フィルタリングシステムの機器構成を示す。図1に示したように、本実施形態の情報フィルタリングシステムは、オペレーティングシステムやユーティリティを含む各種

アプリケーションプログラム（フィルタリング処理を行なう各種プログラムもこれらに含まれる）を実行制御するＣＰＵ1、アプリケーションプログラムや各種データを格納する記憶装置2、および他の計算機からデータを読み込むための回線入出力装置3からなる。なお、この発明は、ソフトウェアとしての実施も可能であり、フロッピィディスクやＣＤ－ＲＯＭなどに格納した形態で提供したり、磁気ディスクなどに格納しておいてネットワークで入手可能な形態で提供することが可能である。

【００２４】図2に本実施形態の情報フィルタリングシステムの機能ブロックを示す。図2に示すように、本実施形態の情報フィルタリングシステムは、制御部１１、新規情報判定部１２、書式解析部１３、複数情報判定部１４、情報分割部１５、類似度算出部１６および結果整形部１７の各処理部を具備してなる。ここでは、これらの各処理部は、ＣＰＵ1で実行制御されるアプリケーションプログラムとして構成されるものとする。

【００２５】制御部１１は、システム全体の動作を制御する。新規情報判定部１２は、処理対象とするデータが新規に発生した情報かどうかを判定する。書式解析部１３は、データ（ＨＴＭＬファイル）の論理的な構造を解析する。複数情報判定部１４は、取り出したデータが複数の内容からなっているかどうかを判定する。

【００２６】また、情報分割部１５は、取り出したデータが複数の内容からなっている場合に、その内容ごとに分割する。類似度算出部１６は、計算対象のデータとプロファイル１６１とを比較して類似度を算出する。結果整形部１５は、類似度の高いデータから順に並べて整形する。

【００２７】ここで、制御部１１の処理の流れを図3を参照して説明する。制御部１１は、監視ページリストに登録されているすべてのページに対して処理を行なう。まず始めに、制御部１１は、監視ページリストからＷｅｂページのアドレスを取り出す（ステップＡ１）。次に、制御部１１は、その取り出したアドレスに基づいて、新規情報判定部１２を実行し（ステップＡ２）、そのアドレスのページが新規情報であるか否かを判定する（ステップＡ３）。新規情報であった場合には（ステップＡ３のＹ）、制御部１１は、書式解析部１３を実行し（ステップＡ４）、対象とするページを取り込むとともに、そのページの論理構造を解析する。次に、制御部１１は、複数情報判定部１４を実行し（ステップＡ５）、処理対象のページが複数の情報単位からなっているページか否かを判定し（ステップＡ６）、複数の情報単位からなっているページである場合には（ステップＡ６のＹ）、情報分割部１５を実行して（ステップＡ７）、このページの内容を各情報単位に分割する。類似度算出部１６では、情報分割部１５で対象ページが分割された場合には、この分割された情報単位ごとに、一方、分割されなかった場合には、そのページ全体を対象に登録されているプロファイル１６１との類似度を算出する（ステップＡ８）。そして、制御部１１は、この算出された類似度を、算出対象の情報単位とともに格納する（ステップＡ９）。

【００２８】監視ページリスト内に処理すべきページアドレスが残っている場合（ステップＡ１０のＹ）、制御部１１は、その残りを対象に始めの処理に戻るが、一方、残りのページが存在しない場合には（ステップＡ１０のＮ）、制御部１１は、結果整形部１７を実行する（ステップＡ１１）。そして、結果整形部１７は、格納されている類似度算出結果を参照し、類似度の高い順に情報単位をソーティングするとともに、利用者に提示する情報フィルタリング結果を生成する。

【００２９】監視ページリストは、システムが監視すべきアドレスの一覧である。利用者がこの監視ページリストに監視したいページアドレスを登録する。

【００３０】次に、新規情報判定部１２の処理の流れを図4を参照して説明する。本実施例では、今回のフィルタリング時に取り込んだページを（ステップＢ１）、前回のフィルタリング時に取り込んだページと比較することにより（ステップＢ２）、そのページに修正が施されたか否かを判定する（ステップＢ３）。変化があった場合（ステップＢ３のＹ）、取り込んだページを次回のフィルタリングに利用するために記憶して（ステップＢ４）、この処理を終了する。なお、ページの作成日や修正日が取り出せる場合には、その情報を用いても良いことはいうまでもない。また、第２実施形態において、Ｗｅｂページの階層関係に対応した新規情報判定処理について述べる。

【００３１】書式解析部１３では、ＨＴＭＬ形式のデータに付与されている各タグに基づいて、Ｗｅｂページの情報を内部構造に変換する。ＨＴＭＬは、ＳＧＭＬのサブセットであり、一般に、開始タグと終了タグとによって論理的な構造を規定している。たとえば、ＨＴＭＬでは、開始タグ＜ＴＩＴＬＥ＞と終了タグ＜／ＴＩＴＬＥ＞とに囲まれた部分がタイトル、および、開始タグ＜ＵＬ＞と終了タグ＜／ＵＬ＞とに囲まれた部分が箇条書きと定義されている。また、段落を規定する＜Ｐ＞や、箇条書きの各項目を表現する＜ＬＩ＞のように、終了タグを省略してよいタグも存在する。これらのタグについては、同じ開始タグが出現した時点で終了タグが存在したものと見なされる。書式解析では、入力データの文字列をスキャンしてＨＴＭＬの開始タグを検出する。そして、その開始タグに対応する終了タグを検出することにより、各タグに対応する情報を取り出す。

【００３２】次に、複数情報判定部１４の処理の流れを図5を参照して説明する。複数情報判定部１４は、箇条書きのフィールドが存在し（ステップＣ１のＹ）、その箇条書きフィールドの各項目に地の文が存在するときに（ステップＣ２のＹ）、各項目の地の文の文字列の平均

長（M）と標準偏差（S）とを求める（ステップC3）。そして、その平均長（M）が、予め定められた長さ（M_0）よりも長く、かつその標準偏差（S）が、予め定められた値（S_0）よりも小さいときに（ステップC4のY）、判定対象のページが、複数の情報単位からなると判定する（ステップC5）。

【0033】図6には、複数の情報単位からなるページのHTMLの記述例、および図7には、そのページの表示イメージが示されている。

【0034】箇条書きの各項目の見出し行は、タグ<LI>と改行タグ
とで囲まれている文字列である。一方、地の文は、見出し行の終わる
から次の<LI>までである。地の文の長さを求めるにあたっては、タグは除外して算出するものとする。HTMLでは、箇条書きのフィールドを定義するタグ<DL>が存在する。<DT>が各項目の見出し行を、<DD>が地の文を表現するタグである。この場合、<DD>から次の<DT>までを地の文として文字列長の計算に用いる。

【0035】なお、処理対象とするページが複数の情報単位からなるページであるかどうかをページごとに記憶する手段を設け、それにしたがって複数の情報単位からなることを判定するようにしても構わない。

【0036】情報分割部15では、箇条書きのフィールドを、各項目ごとに分割して出力する。具体的には、複数情報判定部14で検出した箇条書きの情報単位（見出しと地の文）に分割する。この分割結果は、図8に示したように、見出し（<HEADING>と</HEADING>とで囲んだ部分）と、地の文（<BODY>と</BODY>とで囲んだ部分からなるデータに変換される。

【0037】類似度算出部16の処理は、たとえばプロファイル161に格納された検索条件と処理対象となる各情報単位とをそれぞれ単語頻度のベクトルとして表現し、これらベクトル間の内積をとることによって類似度を求めるといった従前の算出方法を流用すればよい。

【0038】次に、結果整形部17の処理の流れを図9を参照して説明する。結果整形部17は、類似度算出部16での類似度算出の対象となった各情報単位を1つの単位として、類似度の値に基づいてソーティングを行なう（ステップD1）。そして、結果整形部17は、このソーティング結果の順に、情報単位の見出しを箇条書きの項目とし（ステップD2）、地の文から要約を生成して出力する（ステップD3）。要約の生成としては、たとえば、ページの前方から数文を取り出すといった簡単な処理でも構わない。図10に、結果整形部17による整形結果の例を示す。2つの情報が抽出された例である。

【0039】本実施形態では、MosaicなどのHTMLブラウザで表示することを想定しているため、HTML形式で整形結果を出力している。これは、フィルタリング結果で選択された文書のオリジナルをアクセスする場合に、その文書形式との統一性を図るためである。したがって、必ずしもこれに限定するものでなく、特殊なブラウザで取り込める形式のデータに変換するように変形することは，ごく容易である。

【0040】このように、本実施形態の情報フィルタリングシステムによれば、単一の内容からなるWebページと、複数の内容からなるWebページとに対し、これらを同時にフィルタリング対象とし、かつ内容に応じた高精度のフィルタリングを可能とすることができる。

【0041】（第2実施形態）次に、第2の実施形態を説明する。前述した第1の実施形態では、監視するページをすべて事前に登録しておく形態について説明した。しかしながら、Webページは、ハイパーテキストにより階層構造を形成することが可能であるため、単一のページだけを登録する形態では問題が生じる場合がある。

【0042】たとえば、図11（a）に示すように、ページ0から参照されている他のページが階層的に関連づけられて存在しており、しかも（b）に示すように、ページ0が個々のページへのリンク情報だけからなっている場合を考える。この場合、新規情報は、新着情報が記載されているページ21や、プレスリリースを記載したページ24に格納されることになるため、ページ0の内容は、ほとんど修正されることがないことは明らかである。したがって、第1実施形態に示したように、監視ページとしてページ0を事前に登録しておいても、新着情報のページ21の情報が更新された際に、その旨を検出することができない。

【0043】本実施形態では、このような問題に対処するため、監視情報を指定するための監視情報指定手段を設ける。そして、利用者は、新規情報の発生を監視する階層の範囲を事前に設定する。一方、新規情報判定部14では、設定された範囲だけ階層の深さをたどり、新規情報か否かを判定する。

【0044】監視情報指定手段では、監視ページリストを図12に示す形式とし、利用者は、監視するページアドレスとそのページから張られたリンクをたどる段数とを設定する（個々のページについてたどる段数を設定するのではなく、すべてのページに関して同じ段数を設定する形態に変形することも可能である）。

【0045】本実施形態における新規情報判定部14の処理の流れを図13に示す。サブルーチンcheckNew（図13（b））は、設定したページの下位層に位置するページが、新規情報を含むかどうかを再帰的にチェックする。前回のフィルタリング時のページと変化があったかどうかは、第1実施形態と同様に、前回のフィルタリング時に取り込んだページの内容と比較することにより実現することができる。

【0046】情報分割部15についても、階層構造をた

どり個々のページごとに情報単位の分割を行なう。第1実施形態の処理を再帰的に実行することにより実現できるので、処理手続きについては説明を省略する。本実施形態における新規情報判定部14では、内容が変化したページを検出した時点で監視ページとして設定したページ以下の階層に変化があったこととし、それ以下のページをたどらない。この場合には、情報分割部15では、監視ページ以下の全ページについて、処理を行なう必要がある。また、新規情報判定部14において、内容の変化したページを検出した以降も、それ以下のページについて変化があったか否かをチェックするようにしてもよい。この場合、情報分割部15は、変化のあったページについてのみ情報分割処理を行なえばよい。

【0047】本実施形態は、比較的小規模なシステムを想定し、システムに監視させるページのアドレスを、監視ページのリストに利用者自らが登録する形態について説明した。一方、大規模なシステムである場合、事前に監視するページのすべてを事前に登録することは困難である。そこで、取り込んだページに記述されているアドレスを順次たどっていくことが考えられる。大規模システムとして実施する場合は、この形態によって取り込むページの範囲を拡大することも可能である。なた、Ｗｅｂページでは、外部のページへリンクを張っている場合がある。このような外部へのリンクについては無視するように変形することも可能である。

【0048】このように本実施形態の情報フィルタリングシステムによれば、階層的なＷｅｂページを監視可能とし、指定した範囲内に新規または修正された情報があるときに、それをもれなく検知することを可能とする。

【0049】（第3実施形態）次に、第3の実施形態を説明する。本実施形態では、他の情報フィルタリング装置が出力する結果とのマージ機能を持つシステムについて説明する。第1および第2の実施形態では、フィルタリング対象とするページが、ＨＴＴＰ手順にしたがって取り込めることを前提としている。一方、利用者が入手したい情報にはＷｅｂページとして公開されていない情報も存在する。

【0050】図14に、他のフィルタリング装置のフィルタリング結果を取り込む動作原理を示す。（ａ）は、あるＷｅｂサーバ30が設定されており、他の情報フィルタ40が、そのＷｅｂサーバ30のＷｅｂページ31に、フィルタリング結果を書き込む。そして、このＷｅｂページ31を本発明における監視ページリスト20に設定しておくことによって、他のＷｅｂページと同様にフィルタリングを行なうことが可能となる。

【0051】一方、（ｂ）は、電子メールやｆｔｐ手順にしたがって、ローカルなネットワークでアクセス可能なファイルとして格納される場合を示している。この場合、取り込まれたファイルの形式にしたがって、情報取得ゲートウェイ60を設けることにより、他のＷｅｂページと同様にフィルタリングを行なえる。他の情報フィルタ40が、ｆｔｐ手順にしたがってフィルタリング結果を出力する場合、予め定められた名前のファイル（ローカルデータベース50内）に情報フィルタ40のフィルタリング結果が書き込まれる。情報取得ゲートウェイ60は、このファイルをＨＴＭＬ形式に変換し、予め定められた名前のファイルに出力する。そして、監視ページリスト20にこのＨＴＭＬファイルを登録しておくことによって、他のＷｅｂページと同様にフィルタリングを行なうことが可能となる。

【0052】電子メールで送付される場合、電子メールは（メールボックスと呼ばれる）電子メール特定のファイルに格納される。他の一般の電子メールとの区別を行なうため、電子メールのＳｕｂｊｅｃｔ欄に予め取り決めた文字列が設定され、他の情報フィルタ40より送信される。情報取得ゲートウェイ60は、予め取り決めた文字列がＳｕｂｊｅｃｔ欄に設定されているメールをＨＴＭＬ形式に変換し、ＨＴＭＬファイルを更新すればよい。

【0053】情報取得ゲートウェイ60の処理の流れは、他の情報フィルタ40の出力するファイルまたは電子メールの形式に依存する。たとえば、図15に例示したデータが配信される場合には、図16に示す手順でＨＴＭＬに変換できる。

【0054】すなわち、入力ファイルを入力バッファに読み込んだ後（ステップG1）、リンクデータのみからなるＨＴＭＬファイル（ファイル0）を初期化する（生成した後、図17に示す文字列を書き込む）（ステップG2）。

【0055】次に、ファイル番号を1に設定し（ステップG3）、入力バッファの先頭より処理を開始し（ステップG4）、ポインタP以降に見出し行があるかをチェックする（ステップG5）。処理対象の入力ファイルでは、行の先頭が「＊」である行が見出しであるので、それを取り出し、ファイル0にその情報と、ファイル番号に相当するファイル名（ファイル番号が1の場合、”１．ｈｔｍ１”）の情報を出力する（ステップG6）。次に、このファイル名に、見出し行と、入力バッファで見出し行に続く地の文とを書き込み（ステップG7）、ファイル番号を1進めて（ステップG8）、繰り返し処理を行なう。

【0056】そして、処理すべき見出しが入力バッファ中に見出せなくなった時点で（ステップG5のN）、ファイル0に、図18に示す文字列を出力して、この処理を終了する。図19には、図15で示したデータを変換した結果が示されている。

【0057】本実施形態では、処理のモジュラリティを高めるため、一旦ＨＴＭＬファイルに変換する実施形態について説明した。モジュラリティを無視すれば、他の情報フィルタリング装置が出力するフィルタリング結果

のファイルを、直接本発明の装置の入力とするように変形することはごく容易である。

【００５８】このように、本本実施形態の情報フィルタリングシステムによれば、他の情報フィルタリング装置が出力したフィルタリング結果を読み込むことにより、単独の情報フィルタリング装置が監視できる以上の範囲の情報を監視することが可能となる。

【００５９】

【発明の効果】以上詳述したように、この発明によれば、複数の形態を有するＷｅｂページをはじめとする文書情報のフィルタリングを統一的に処理し、利用者の分かりやすい形態で提供することができる。

【００６０】第１の発明にあっては、複数の情報単位からなる文書内の各情報単位について、回りのテキストに影響されることなく独立して類似度を算出するため、高い精度でフィルタリング処理を行なうことが可能となる。

【００６１】また、第２の発明にあっては、ハイパーテキスト形式の文書を、フィルタリング対象とする際階層の段数を指定することにより、複数のＷｅｂページで一つの情報を表現しているＷｅｂページ群を効果的に更新監視させることができ、また、無制限に階層をたどることを排除することができるため、処理時間を抑えることが可能となる。

【００６２】さらに、第３の発明にあっては、他の情報フィルタリング装置の出力結果を、他の文書と同じようにマージして出力でき、利用者に分かりやすい結果を提供することが可能となる。

【図面の簡単な説明】

【図１】第１実施形態の情報フィルタリングシステムの機器構成を示す図。

【図２】同実施形態の情報フィルタリングシステムの機能ブロックを示す図。

【図３】同実施形態の制御部の処理の流れを示すフローチャート。

【図４】同実施形態の新規情報判定部の処理の流れを示すフローチャート。

【図５】同実施形態の複数情報判定部の処理の流れを示すフローチャート。

【図６】同実施形態の複数の情報単位からなるページのＨＴＭＬの記述例を示す図。

【図７】図６で示したＨＴＭＬ記述の表示イメージを示す図。

【図８】同実施形態の情報分割部の分割結果を示す図。

【図９】同実施形態の結果整形部の処理の流れを示すフローチャート。

【図１０】同実施形態の結果整形部の整形結果の例を示す図。

【図１１】ハイパーテキストによって階層構造を形成するＷｅｂページを説明する図。

【図１２】第２実施形態の監視ページリストの形式を示す図。

【図１３】同実施形態の新規情報判定部の処理の流れを示すフローチャート。

【図１４】第３実施形態の他のフィルタリング装置のフィルタリング結果を取り込む動作原理を示す図。

【図１５】同実施形態の配信されるデータを例示する図。

【図１６】同実施形態の配信データをＨＴＭＬに変換する手順を示すフローチャート。

【図１７】同実施形態のＨＴＭＬファイルに書き込まれる記述を示す図。

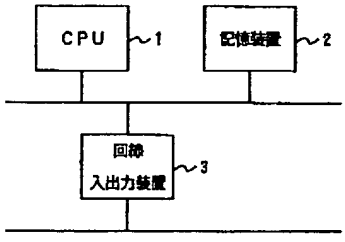
【図１８】同実施形態のＨＴＭＬファイルに書き込まれる記述を示す図。

【図１９】図１５で示した配信データをＨＴＭＬに変換した結果を示す図。

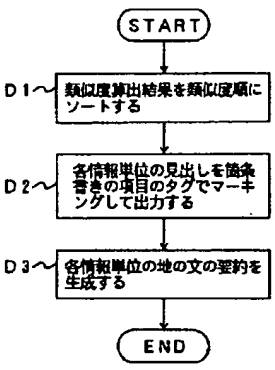
【符号の説明】

１…ＣＰＵ、２…記憶装置、３…回線入出力装置、１１…制御部、１２新規情報判定部、１３…書式解析部、１４…複数情報判定部、１５…情報分割部、１６…類似度算出部、１６１…プロファイル、１７…結果整形部。

【図１】



【図９】



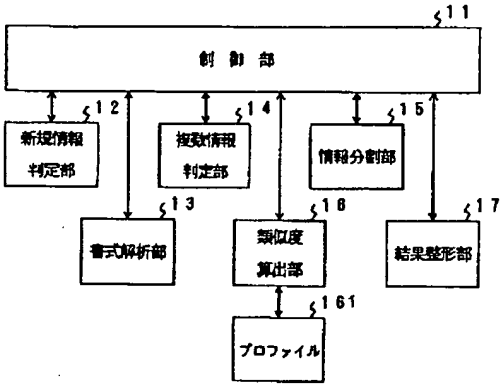
【図１２】

アドレス	階層数
アドレス１	1
アドレス２	2
アドレス３	3
｜	｜

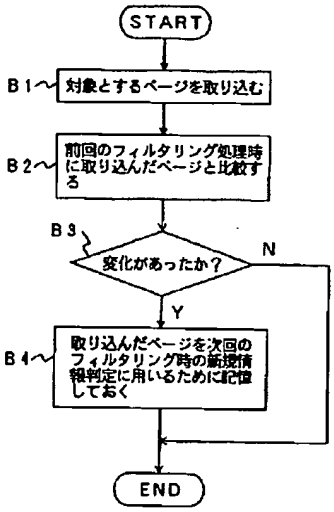
【図１８】

```
</UL>
</BODY>
</HTML>
```

【図2】



【図4】



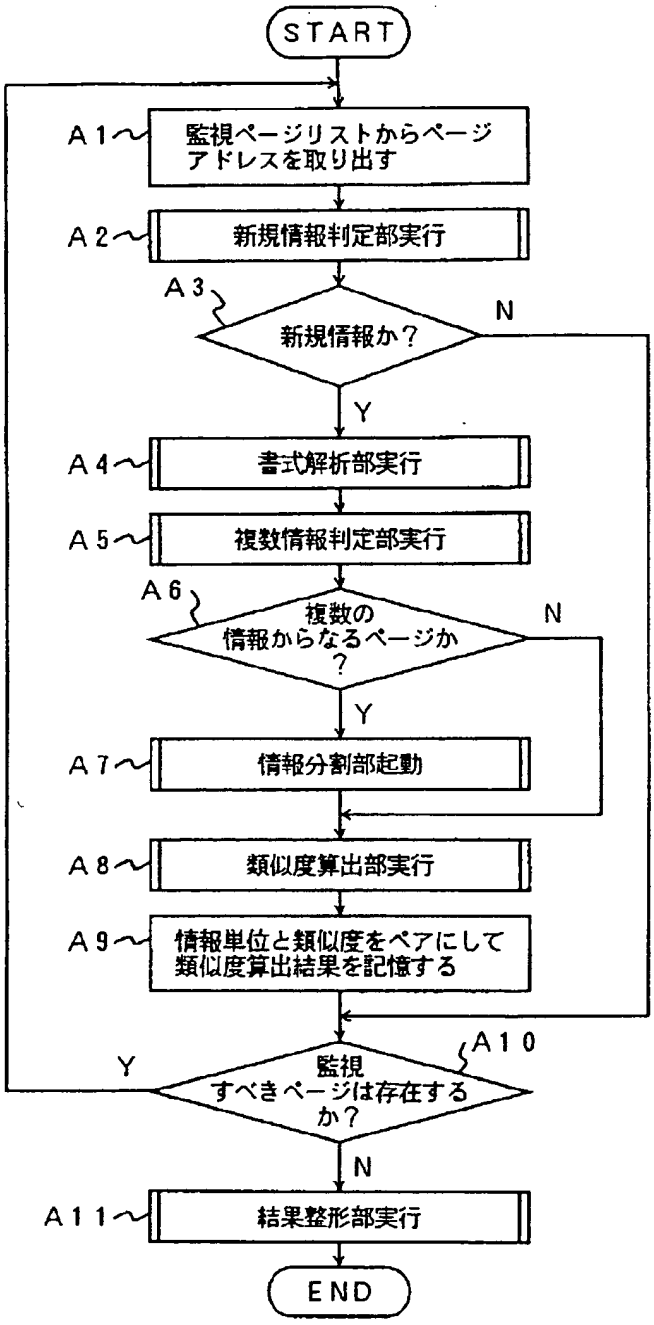
【図8】

<HEADING><A HREF="http://www.aaaaaa.co.jp/press.html">○○商事・○○物産、全国に××流通網――系列卸通じ集荷・販売</A><HEADING>
<BODY>
○○商事と○○物産は系列の大手食品卸を通じて国内の主食用××流通に本格参入する。十一月の新食糧法（食糧需給価格安定法）施行で××市場の規制が緩和されるのをにらみ、集荷から販売まで一貫して行う。総合商社では○○商事などが小売業者などに資本参加した例はあるが、全国規模で××の集荷・販売に踏み切るのは初めて。××市場は四兆円といわれる。両社合わせて当面、千億円規模の売り上げを目指すとみられる。○○○○協同組合連合会（○○）など○○グループや既存の××卸業者などの戦略にも大きな影響を与えそうだ。
</BODY>

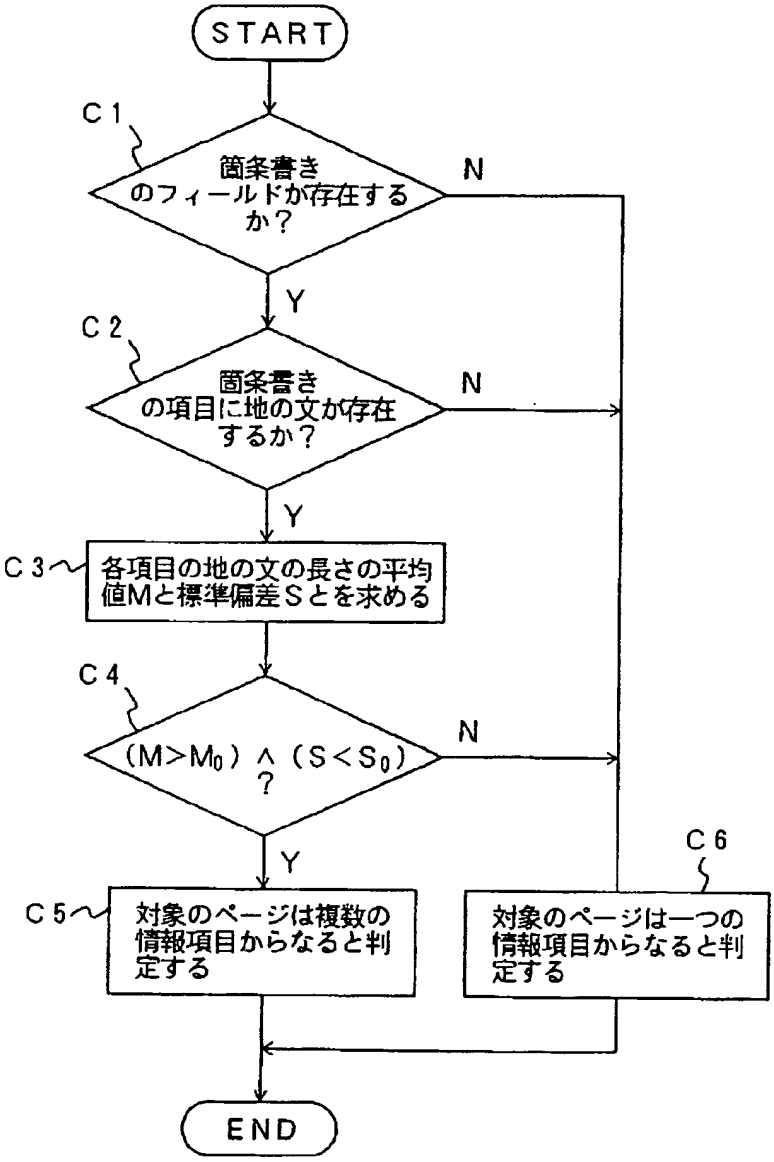
【図17】

```
<HTML>
<!- ファイル0 ->
<HEAD>
<TITLE>Filtering result</TITLE>
</HEAD>
<BODY>
<UL>
```

【図3】


START
A1
監視ページリストからページアドレスを取り出す
A2
新規情報判定部実行
A3
新規情報か？
N
Y
A4
書式解析部実行
A5
複数情報判定部実行
A6
複数の情報からなるページか？
N
Y
A7
情報分割部起動
A8
類似度算出部実行
A9
情報単位と類似度をペアにして類似度算出結果を記憶する
A10
監視すべきページは存在するか？
Y
N
A11
結果整形部実行
END

【図5】


START
C1
箇条書きのフィールドが存在するか？
N
Y
C2
箇条書きの項目に地の文が存在するか？
N
Y
C3
各項目の地の文の長さの平均値Mと標準偏差Sとを求める
C4
(M>M0) ∧ (S<S0) ？
N
Y
C5
対象のページは複数の情報項目からなると判定する
C6
対象のページは一つの情報項目からなると判定する
END

【図6】

```
<HTML>
<HEAD>
<TITLE>最新情報インデックス</TITLE>
</HEAD>
<BODY>
<UL>
<LI><A HREF="http://www.aaaaaa.co.jp/press.html">○○商事・○○物産、全国に××流通網――系列卸通じ集荷・販売</A><BR>
○○商事と○○物産は系列の大手食品卸を通じて国内の主食用××流通に本格参入する。十一月の新食糧法（食糧需給価格安定法）施行で××市場の規制が緩和されるのをにらみ、集荷から販売まで一貫して行う。総合商社では○○商事などが小売業者などに資本参加した例はあるが、全国規模で××の集荷・販売に踏み切るのは初めて。××市場は四兆円といわれる。両社合わせて当面、千億円規模の売り上げを目指すとみられる。○○○○協同組合連合会（○○）など○○グループや既存の××卸業者などの戦略にも大きな影響を与えそうだ。
<BR><HR>

<LI><A HREF="http://www.bbbbb.co.jp/index.html">○○○○出版子会社、××に責任販売制――返品制限、マージン高く</A><BR>
○○○○○○○の出版子会社、○○○○出版（東京・港、○○○○社長）は、十月に生活情報誌を創刊するのを機に「責任販売制」を導入する。出版流通は「返品自由」が原則だが、この「責任販売制」は、書店に返品率の上限を設定する代わりに、書店側の仕入れ部数に応じてマージン（利幅）を引き上げる仕組み。一般書店を対象とした××流通では初の試みで、返品による物流経費の無駄を省き、「書店の生産性を高める方向で、実需に応じた効率的な流通の仕組みを作る」（○○社長）のが狙い。<BR><HR>

<LI><A HREF="http://www.cccccc.co.jp/press.html"> 簡易双方向ＴＶ、ゲーム機感覚で番組参加――○○・○○など今秋開始</A><BR>
○○など電機メーカーと○○○○○○○○などの民放、○○○○○○○○○○が協力して、今秋にも簡易型双方向テレビの放送を始める。専用チューナーを取り付けたテレビを電話回線につなぎ、テレビゲーム機のような簡単なコントローラーを操作して番組に参加する仕組み。テレビショッピングで画面に出ている商品をその場で申し込んだり、クイズ番組を見ながら直接回答することができる。各種世論調査など応用範囲が広く、マルチメディア放送が本格化するきっかけになりそうだ。

</UL>
</BODY>
</HTML>
```

【図７】

最新情報インデックス
□ ○○商事・○○物産、全国に××流通網――系列卸通じ集荷・販売 ○○商事と○○物産は系列の大手商品卸を通じて国内の主食用××流通に本格参入する。十一月の新食糧法（食糧需給価格安定法）施行で××市場の規制が緩和されるのをにらみ、集荷から販売まで一貫して行う。総合商社では○○商事などが小売業者などに資本参加した例はあるが、全国規模で××の集荷・販売に踏み切るのは初めて。××市場は四兆円といわれる。両社合わせて当面、千億円規模の売り上げを目指すとみられる。○○○○協同組合連合会（○○）など○○グループや既存の××卸業者などの戦略にも大きな影響を与えそうだ。
□ ○○○○出版子会社、××に責任販売制――返品制限、マージン高く ○○○○○○○の出版子会社、○○○○出版（東京・港、○○○○社長）は、十月に生活情報誌を創刊するのを機に「責任販売制」を導入する。出版流通は「返品自由」が原則だが、この「責任販売制」は、書店に返品率の上限を設定する代わりに、書店側の仕入れ部数に応じてマージン（利幅）を引き上げる仕組み。一般書店を対象とした××流通では初の試みで、返品による物流経費の無駄を省き、「書店の生産性を高める方向で、実需に応じた効率的な流通の仕組みを作る」（○○社長）のが狙い。
□ 簡易双方向ＴＶ、ゲーム機感覚で番組参加――○○・○○など今秋開始 ○○など電機メーカーと○○○○○○○○などの民放、○○○○○○○○○○○○が協力して、今秋にも簡易型双方向テレビの放送を始める。専用チューナーを取り付けたテレビを電話回線につなぎ、テレビゲーム機のような簡単なコントローラーを操作して番組に参加する仕組み。テレビショッピングで画面に出ている商品をその場で申し込んだり、クイズ番組を見ながら直接回答することができる。各種世論調査など応用範囲が広く、マルチメディア放送が本格化するきっかけになりそうだ。

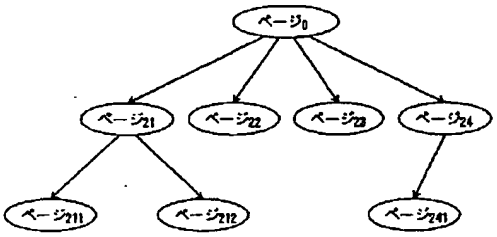
【図10】

```
<HTML>
<HEAD>
<TITLE>フィルタリング結果</TITLE>
</HEAD>
<BODY>
<UL>
<LI><A HREF="http://www.ccccccc.co.jp/press.html.>簡易双方向ＴＶ、ゲー
ム機感覚で番組参加－－○○・○○など今秋開始</A><BR>
○○など電機メーカーと○○○○○○○○などの民放、○○○○○○○○○○○
が協力して、今秋にも簡易型双方向テレビの放送を始める。

<LI><A HREF="http://www.ddd.or.jp/press.html.>○○○、○○○と広範な協力</A><BR>
○○○は二十日、○○○○○○○○○○○と広範な分野で協力することで合意
した。同日、○○○の○○○○会長が○○○○○○○○○○会長と会談、「協
力協定」を締結した。

</UL>
</BODY>
</HTML>
```

【図11】

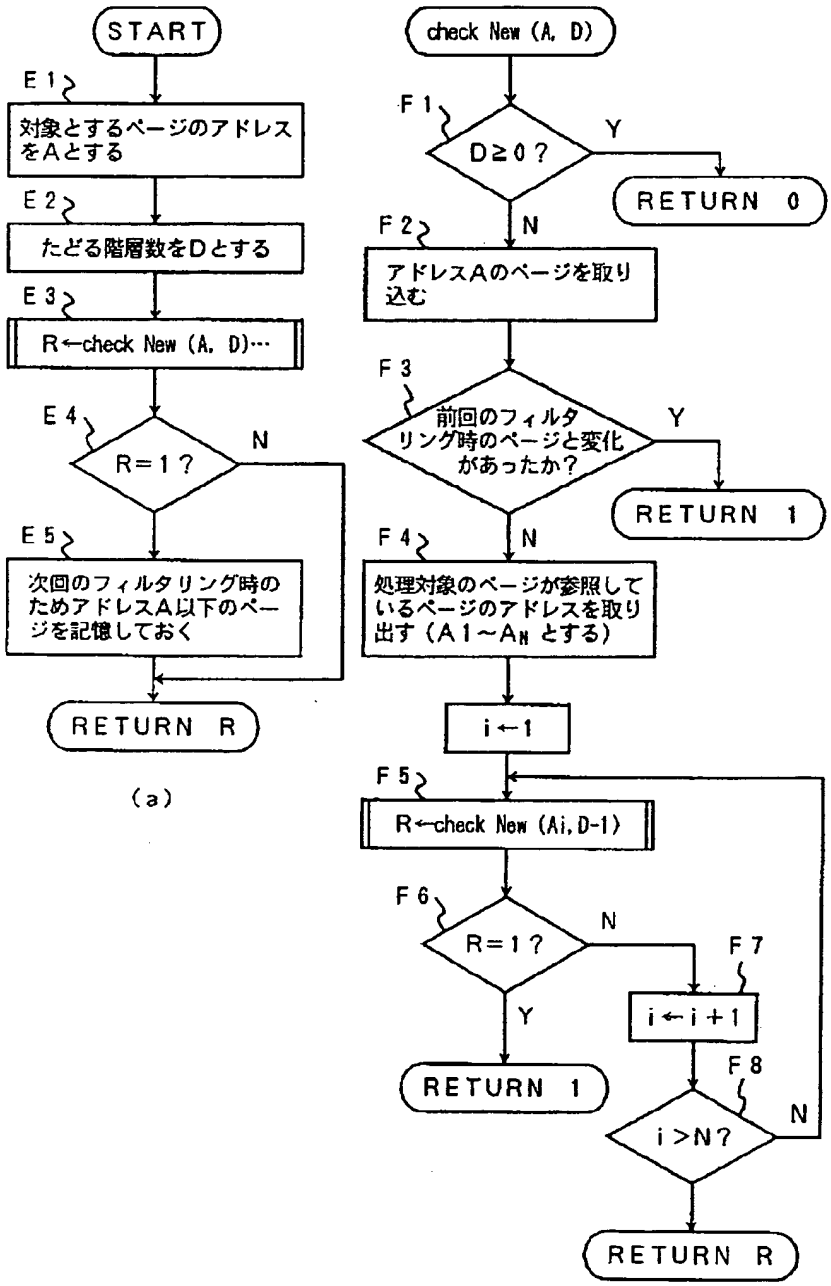


(a)

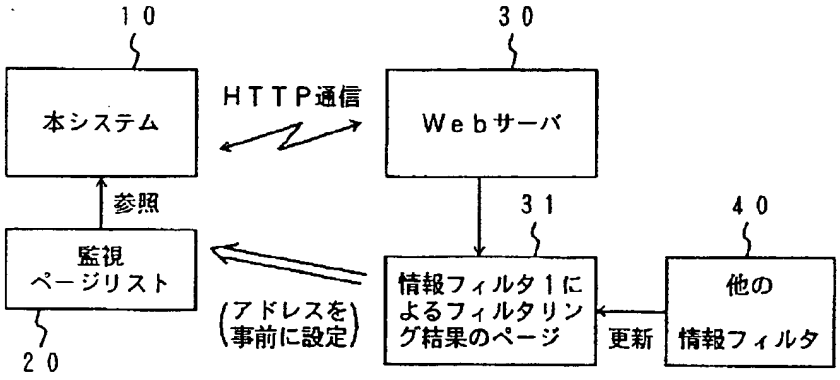
```
<UL>
  <LI><A HREF="page21.html"> 新着情報</A>
  <LI><A HREF="page22.html"> フィードバック</A>
  <LI><A HREF="page23.html"> ページの概要</A>
  <LI><A HREF="page24.html"> プレスリリース</A>
</UL>
```

(b)

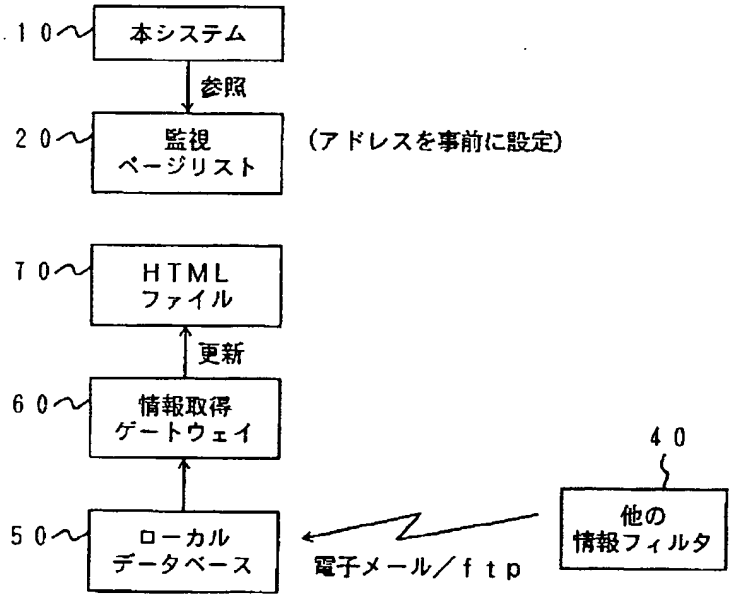
【図13】


START
E1
対象とするページのアドレスをAとする
E2
たどる階層数をDとする
E3
R←check New (A, D)…
E4
R=1？
N
E5
次回のフィルタリング時のためアドレスA以下のページを記憶しておく
RETURN R
(a)
check New (A, D)
F1
D≧0？
Y
RETURN 0
N
F2
アドレスAのページを取り込む
F3
前回のフィルタリング時のページと変化があったか？
Y
RETURN 1
N
F4
処理対象のページが参照しているページのアドレスを取り出す（A1～AN とする）
i←1
F5
R←check New (Ai, D-1)
F6
R=1？
N
F7
i←i+1
Y
RETURN 1
F8
i>N？
N
RETURN R
(b)

【図14】


10
30
本システム
HTTP通信
Webサーバ
参照
31
40
監視
ページリスト
情報フィルタ1に
よるフィルタリン
グ結果のページ
他の
情報フィルタ
更新
(アドレスを
事前に設定)
20

(a)


10
本システム
参照
20
監視
ページリスト
(アドレスを事前に設定)
70
HTML
ファイル
更新
60
情報取得
ゲートウェイ
40
50
ローカル
データベース
他の
情報フィルタ
電子メール／ftp

(b)

【図15】

＊独○○○○、××××通販品目を拡充－－○○○と資本提携（AA新聞、朝刊、2面）
○○○の通販大手、○○○○（○○○○○）は、××××関連製品卸売業者を傘下に持つ△△△△△△△△△△△△△△△△△△△△△△（△△△△△）と資本提携した。○○○○は△△△から××××や周辺製品を安く調達、カタログ商品の内容を大幅に充実する。欧州でも家庭での××××需要が急激に拡大しているが、○○○○はこれに対応する。

同社は一月一日にさかのぼって△△△に四五％資本参加する。△△△の子会社で××××関連製品卸売会社、□□□□□の昨年の売り上げは前年比四○％強増の十六億マルク（一マルク＝約六○円）で、欧州市場で二○％強のシェアを獲得し、同市場で第二位、全世界では第五位の大手。

○○○○によると、これまでも××××関連製品を各メーカーから直接調達してカタログ商品に取り入れていたが、多数のメーカーの製品を卸す□□□□□からの方が調達条件が良い上、品ぞろえも容易になるメリットがある。

○○○では××××の家庭への普及率は米国に比べかなり低い。○○○○は、マルチメディアブームやパソコン通信ユーザーの急増に加え、××××の本体の急速な低価格化と高性能化も手伝って、家庭向け需要が今後急速に高まるとみている。（ボン＝宮尾猛）

＊○○○○社長が退任　経営方針で親会社と相違（BB通信）

○○○○○○○○○○（東京本社）は３日、○○○○社長が６月２８日付で退任したと発表した。○○氏は親会社の米○○○○○○○○○○の副社長も辞任した。理由について同社は「経営方針に対する考え方の相違があった」と説明している。

米○○○○○○○○○○の上級副社長で日本法人会長の○○○○○○○○○○氏が同日付で東京常駐となり、日本での事業を統括する。同時に後任社長の人選を進める。

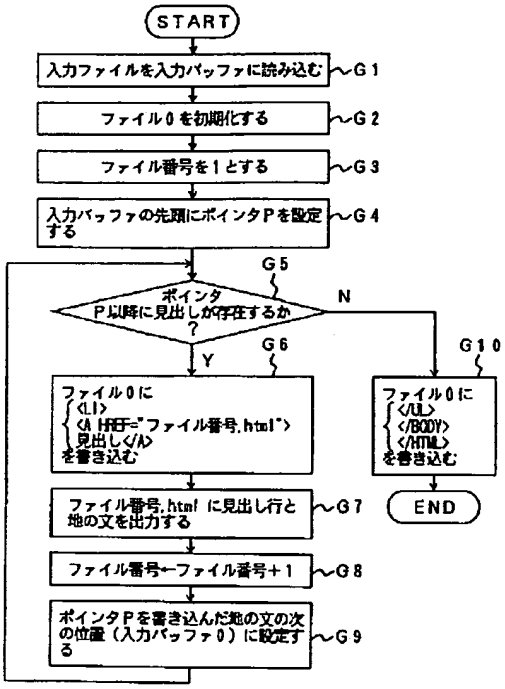
＊○ＴＶ大手○○○○、××××を販売（AA新聞、朝刊、２面）

【ボン＝宮尾猛】○○○のテレビ放送大手、○○○○が国内の中堅××××メーカーと組み、九月からマルチメディア対応型××××を売り出す。高性能、簡単操作、低価格を武器に家庭ユーザーの獲得を目指す。すでに××××チェーン大手の○○○○や○○○○○○○○などの百貨店と連携、国内全域に千六百カ所の販売拠点を確保した。

○○○○と組むのは△△△△△△△△△△△△△△△△△△△△△（△△△△△）。○○○○は○○○の生産する同対応型××××を「○○○○・ファミリーＰＣ」のブランドで売る。○○○○○のＣＰＵ（中央演算処理装置）「××××××」を搭載し、ＣＤ－ＲＯＭ（コンパクトディスクを使った読み出し専用メモリー）駆動装置、サウンドカード、スピーカーを内蔵する。

価格はモニター、キーボード、マウス込みで入門クラスが二千四百マルク（一マルク＝約六一円）、上級クラスは三千マルク前後を予定しており、「低価格で定評のある国内最大手○○○○○とも十分競走できる」（○○○○○○○○○社長）という。

【図16】


START
入力ファイルを入力バッファに読み込む
G1
ファイル0を初期化する
G2
ファイル番号を1とする
G3
入力バッファの先頭にポインタPを設定する
G4
G5
ポインタP以降に見出しが存在するか？
N
Y
G6
G10
ファイル0に
<LI>
<A HREF="ファイル番号.html">
見出し</A>
を書き込む
ファイル0に
</UL>
</BODY>
</HTML>
を書き込む
ファイル番号.html に見出し行と地の文を出力する
G7
END
ファイル番号←ファイル番号＋1
G8
ポインタPを書き込んだ地の文の次の位置（入力バッファ0）に設定する
G9

【図19】

```
<HTML>
<!- ファイル0 ->
<HEAD>
<TITLE>Filtering result</TITLE>
</HEAD>
<BODY>
<UL>
<LI>
<A HREF="1.html">
独○○○○、××××通販品目を拡充－－○○○と資本提携（AA新聞、朝刊、
2面）</A>
<LI>
<A HREF="2.html">
○○○○社長が退任　運営方針で親会社と相違（BB通信）</A>
<LI>
<A HREF="3.html">
○ＴＶ大手○○○○、××××を販売（AA新聞、朝刊、２面）</A>
</UL>
</BODY>
</HTML>
```

（ａ）ファイル0（リンクのみのＨＴＭＬファイル）

```
<HTML>
<!- ファイル1 ->
<HEAD>
<TITLE>result 1</TITLE>
</HEAD>
<BODY>
<H1>○○○○○、××××通販品目を拡充－－○○○と資本提携（AA新聞、朝
刊、２面）</H1>
<P>
○○○の通販大手、○○○○（○○○○○）は、××××関連製品卸売業者を
傘下に持つ△△△△△△△△△△△△△△△△△△△△△△△（○○○○○）と
資本提携した。○○○○は△△△から××××や周辺製品を安く調達、カタロ
グ商品の内容を大幅に充実する。欧州でも家庭での××××需要が急激に拡大
しているが、○○○○はこれに対応する。
<P>
同社は一月一日にさかのぼって△△△に四五％資本参加する。△△△の子会社
で××××関連製品卸売会社、□□□□□の昨年の売り上げは前年比四○％強
増の十六億マルク（一マルク＝約六○円）で、欧州市場で二○％強のシェアを
獲得し、同市場で第二位、全世界では第五位の大手。
<P>
○○○○によると、これまでも××××関連製品を各メーカーから直接調達し
てカタログ商品に取り入れていたが、多数のメーカーの製品を卸す□□□□□
からの方が調達条件が良い上、品ぞろえも容易になるメリットがある。
<P>
○○○では××××の家庭への普及率は米国に比べかなり低い。○○○○は、
マルチメディアブームやパソコン通信ユーザーの急増に加え、××××の本体
の急速な低価格化と高性能化も手伝って、家庭向け需要が今後急速に高まると
みている。（ボン＝宮尾猛）
</BODY>
</HTML>
```

（ｂ）ファイル1（第１番目の記事に対応するＨＴＭＬファイル）